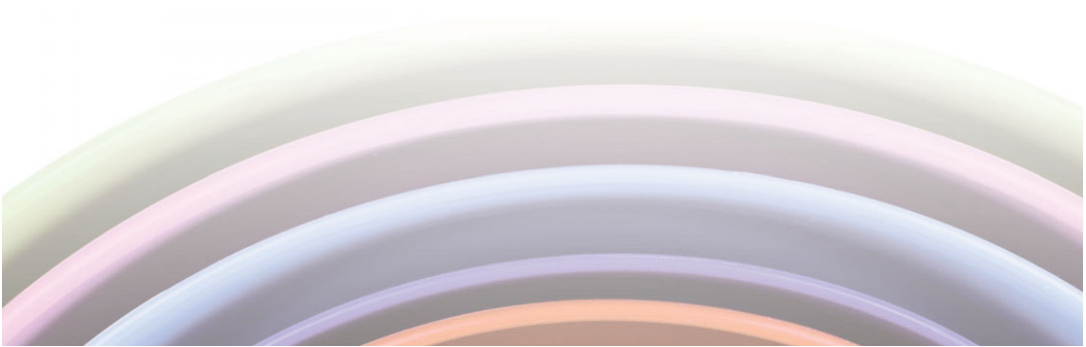


12

Structured observation

Chapter outline

Introduction	270
Problems with survey research on social behaviour	270
So why not observe behaviour?	272
The observation schedule	275
Strategies for observing behaviour	276
Sampling	277
Sampling people	277
Sampling in terms of time	278
Further sampling considerations	278
Issues of reliability and validity	279
Reliability	279
Validity	280
Field stimulations as a form of structured observation	282
Criticisms of structured observation	283
On the other hand . . .	284
Checklist	285
Key points	285
Questions for review	286





Chapter guide

Structured observation is a method that is relatively underused in social research. It entails the direct observation of behaviour and the recording of that behaviour in terms of categories that have been devised prior to the start of data collection. This chapter explores:

- the limitations of survey research for the study of behaviour;
- the different forms of observation in social research;
- the potential of structured observation for the study of behaviour;
- how to devise an observation schedule;
- different strategies for observing behaviour in structured observation;
- sampling issues in structured observation research; with this method, the issue of sampling is to do not just with people but also with the sampling of time and contexts;
- issues of reliability and validity in structured observation;
- field stimulations, whereby the researcher actively intervenes in social life and records what happens as a consequence of the intervention, as a form of structured observation;
- some criticisms of structured observation.

Introduction

Structured observation is a method for systematically observing the behaviour of individuals in terms of a schedule of categories. It is a technique in which the researcher employs explicitly formulated rules for the observation and recording of behaviour. One of its main advantages is that it allows behaviour to be observed directly, unlike in

survey research, which allows behaviour only to be inferred. In survey research, respondents frequently report their behaviour, but there are good reasons for thinking that such reports may not be entirely accurate. Structured observation constitutes a possible solution in that it entails the direct observation of behaviour.



Problems with survey research on social behaviour

Chapters 8 through 11 have dealt with several different aspects of survey research. In the course of outlining procedures associated with the survey, certain problems with the techniques with which it is typically associated have been identified. To some extent the deficiencies associated with the survey are recognized by practitioners, who have developed ways of dealing with them or at least of offsetting their impact to some degree. When survey techniques such as the structured interview or the

self-completion questionnaire are employed in connection with the study of respondents' *behaviour*, certain characteristic difficulties are encountered, some of which have been touched on in earlier chapters. Tips and skills 'Problems with using survey research to investigate behaviour' identifies some of the difficulties entailed in using survey methods to research behaviour. The list is by no means exhaustive but it does capture some of the main elements.



Tips and skills

Problems with using survey research to investigate behaviour

- *Problem of meaning.* People may vary in their interpretations of key terms in a question (see Thinking deeply 12.1).
- *Problem of omission.* When answering the question, respondents may inadvertently omit key terms in the question (see Thinking deeply 12.1).
- *Problem of memory.* They may misremember aspects of the occurrence of certain forms of behaviour.
- *Social desirability effect.* They may exhibit a tendency towards replying in ways that are meant to be consistent with their perceptions of the desirability of certain kinds of answer.
- *Question threat.* Some questions may appear threatening and result in a failure to provide an honest reply.
- *Interviewer characteristics.* Aspects of the interviewer may influence the answers provided.
- *Gap between stated and actual behaviour.* How people say they are likely to behave and how they actually behave may be inconsistent (see Thinking deeply 12.2).



Thinking deeply 12.1

Accurate reporting of behaviour and the problems of meaning and omission

Belson (1981) has conducted detailed studies of how people interpret questions designed to gauge attitudes and behaviour. One question concerned with the latter was embedded in a structured interview schedule administered to fifty-nine British adults and went as follows:

When you turn on your television in the evening, do you generally go on viewing till the end of the evening or do you just watch one or two programmes? (Belson 1981: 59)

Intensive interviews undertaken after the structured interviews had been carried out revealed that no respondents interpreted the question totally correctly. Twenty-five respondents arrived at incorrect interpretations; the rest were broadly correct but in varying degrees. A common problem was that the question was designed to refer to when the respondents themselves turned the television on. This was correctly interpreted by thirty-eight respondents, but fifteen interpreted the question to mean when the set was switched on—that is, not necessarily by the respondent (problem of meaning). Nine respondents appeared not to have taken any notice of the phrase 'when you turn on your television' (problem of omission). Similarly, ten failed to consider 'till the end of the evening' in their answers, while 'generally' spawned several interpretations. We see here problems of omission and meaning respectively.



Thinking deeply 12.2

Gap between stated and actual behaviour

This is one of the most infamous cases of problems of the gap between what people say they do (or are likely to do) and their actual behaviour. Questionnaires tap people's attitudes and reports of their behaviour, but one might legitimately question how well these relate to actual behaviour. A study of racial prejudice conducted many years ago by LaPiere (1934) illustrates this issue. He spent two years travelling with a young Chinese student and his

wife to determine whether they were refused entry at hotels and restaurants. They twice crossed the USA. Of 66 hotels, they were refused entry once; of 184 restaurants and diners, none refused entry. LaPiere sought to eliminate himself as a possible contaminating influence by ensuring that he was not involved in gaining access to the various establishments and indeed seems to have sought to load the dice slightly in favour of being turned away:

Whenever possible I let my Chinese friend negotiate for accommodation . . . or sent them into a restaurant ahead of me. In this way I attempted to 'factor' myself out. We sometimes patronized high-class establishments after a hard and dusty day on the road and stopped at inferior auto camps when in our most presentable condition. (LaPiere 1934: 232)

LaPiere then allowed six months to elapse and sent questionnaires to the hotels and restaurants they had visited. One of the questions asked: 'Will you accept members of the Chinese race as guests in your establishment?' Of the establishments that replied, 92 per cent of restaurants said no; and 91 per cent of hotels said no. LaPiere's simple though striking study clearly illustrates the gap that may exist between reports of behaviour and actual behaviour. It should also be noted that the question asked is somewhat unclear, a feature that is not usually remarked upon in connection with this widely cited study. 'Will you . . . ?' can be interpreted as asking the respondent to project into the future or to state the establishment's policy. Quite why the more obvious formulation of 'Do you . . . ?' was not used is not clear, though it is unlikely that this point has a significant bearing on the findings and their implications for survey research.



So why not observe behaviour?

An obvious solution to the problems identified is to observe people's behaviour directly rather than to rely on research instruments like questionnaires to elicit such information. In this chapter, I am going to outline a method called *structured observation* (see Key concept 12.1), also often called **systematic observation**.

Much like the interview (see Key concept 9.2), there are many different forms of observation approach in social research. Key concept 12.2 outlines some major ways of conducting observation studies in social research.

It has been implied that structured observation can be viewed as an alternative to survey methods of research. After all, in view of the various problems identified in Tips and skills 'Problems with using survey research to investigate behaviour', it would seem an obvious solution to observe people instead. However, structured observation has not attracted a large following and instead tends to be used in certain specific research areas, such as the behaviour of school teachers and pupils and interaction between them.



Key concept 12.1 What is structured observation?

Structured observation, often also called *systematic observation*, is a technique in which the researcher employs explicitly formulated rules for the observation and recording of behaviour. The rules inform observers about what they should look for and how they should record behaviour. Each person who is part of the research (we will call these people 'participants') is observed for a predetermined period of time using the same rules. These rules are articulated in what is usually referred to as an **observation schedule**, which bears many similarities to a structured interview schedule with closed questions. The aim of the observation schedule is to ensure that each participant's behaviour is systematically recorded so that it is possible to aggregate the behaviour of all those in the sample in respect of each type of behaviour being recorded. The rules that constitute the observation schedule are as specific as possible in order to direct observers to exactly what aspects of behaviour they are supposed to be looking for. The resulting data resemble questionnaire data considerably, in that the procedure generates information on different aspects of behaviour that can be treated as variables. Moreover, structured observation research is typically underpinned by a cross-sectional research design (see Key concept 3.6 and Figures 3.2 and 3.3).



Key concept 12.2

Major types of observation research

The following are the major forms of observational research in social research.

- *Structured observation.* See Key concept 12.1.
- *Systematic observation.* See Key concept 12.1.
- *Participant observation.* This is one of the best-known methods of research in the social sciences. It is primarily associated with qualitative research and entails the relatively prolonged immersion of the observer in a social setting in which he or she seeks to observe the behaviour of members of that setting (group, organization, community, etc.) and to elicit the meanings they attribute to their environment and behaviour. Participant observers vary considerably in how much they participate in the social settings in which they locate themselves. See Key concept 19.1 and Chapter 19 generally for a more detailed treatment.
- *Non-participant observation.* This is a term that is used to describe a situation in which the observer observes but does not participate in what is going on in the social setting. Structured observers are usually non-participants in that they are in the social setting being observed but rarely participate in what is happening. The term can also be used in connection with unstructured observation.
- *Unstructured observation.* As its name implies, unstructured observation does not entail the use of an observation schedule for the recording of behaviour. Instead, the aim is to record in as much detail as possible the behaviour of participants with the aim of developing a narrative account of that behaviour. In a sense, most participant observation is unstructured, but the term unstructured observation is usually employed in conjunction with non-participant observation.
- *Simple observation and contrived observation.* Webb et al. (1966) write about forms of observation in which the observer is unobtrusive and is not observed by those being observed. With simple observation, the observer has no influence over the situation being observed; in the case of contrived observation, the observer actively alters the situation to observe the effects of an intervention. These two types of observation are invariably forms of non-participant observation and can entail either structured or unstructured observation.

Central to any structured observation study will be the *observation schedule* or *coding scheme*. This specifies the categories of behaviour that are to be observed and how behaviour should be allocated to those categories. It is best to illustrate what this involves by looking at examples. One of the best-known schedules for the observation of classrooms is Flanders Interaction Analysis Categories (FIAC), devised by Flanders (1970). This scheme was developed in the USA, but has been employed fairly extensively in other countries. An observer watches a classroom during a lesson and every three seconds allocates a category number to the type of activity that takes place in that three-second period. (See Figure 12.1 for the different types of activity in the FIAC scheme.) In other words, in each and every minute, the observer will write down twenty numbers, each of which will relate directly to the coding scheme.

From such data a number of features can be derived. For example, it becomes possible to compare teachers' styles in terms of such things as the relative emphasis

upon teachers doing the talking and pupils talking. One can also compare teachers in terms of the amounts of silence or confusion that take place in their lessons. It also becomes possible to compare classes in terms of these categories. In fact, we tend to find, using the FIAC scheme, that two-thirds of time in the classroom is made up of talk. Of that talk, two-thirds derives from the teacher. It is helpful in bringing about an understanding of what happens in lessons and can be useful in developing information about which styles seem most effective. For example, are better exam results achieved by teachers who exhibit a high level of talking relative to that of pupils, or are teachers who allow more pupil talk more effective? The scheme can also be used in teacher training in order to help trainees to become aware of features of their teaching style, and possibly to begin to question its appropriateness.

It is interesting to think about how a scheme like this might be employed in connection with higher education teaching and in particular in tutorials and seminars. In

Figure 12.1

FIAC categories

<i>Teacher talk</i>	Response	{ 1 Accepts feeling (e.g. accepts and clarifies an attitude or the feeling tone of a pupil) 2 Praises or encourages 3 Accepts or uses ideas of pupils
	Initiation	{ 4 Asks questions 5 Lecturing 6 Giving direction 7 Criticizing or justifying authority
<i>Pupil talk</i>		8 Pupil talk—response
		9 Pupil talk—initiation
<i>Silence</i>		10 Silence or confusion

the following imaginary scheme, the focus is on the teacher. The categories might be:

Tutor

1. asking question addressed to group;
2. asking question addressed to individual;
3. responding to question asked by member of group;
4. responding to comment by member of group;
5. discussing topic;
6. making arrangements;
7. silence.

Student(s)

8. asking question;
9. responding to question from tutor;
10. responding to comment from tutor;
11. responding to question from another student;
12. responding to comment from another student;
13. talking about arrangements.

We might want to code what is happening every five seconds. The coding sheet for a five-minute period in the tutorial might then look like Figure 12.2. In this grid, each cell represents a five-second interval so that a row comprises twelve five-second intervals—that is, a minute.

Figure 12.2

Coding sheet for imaginary study of university tutors

3	3	3	3	10	10	10	10	10	10	10	10	10
10	10	10	10	10	10	7	7	7	8	8	8	8
8	8	8	8	8	8	8	8	8	11	11	11	11
11	11	11	11	11	11	11	11	11	11	11	11	11
7	7	7	7	7	4	4	4	4	4	4	4	1

Note: Each cell represents a five-second interval and each row is one minute. The number in each cell refers to the code used to represent a category of behaviour that has been observed.

The numbers in each cell are the codes used to represent the classification of behavior. Thus, the top left-hand cell has a 3 in it, which refers to a tutor responding to a question asked by a member of the group. We might try to relate the amount of time that the tutor is engaged in particular activities to such things as: number of students in the group; layout of the room; subject discipline; gender of tutor; age of tutor; and so on.



Research in focus 12.1

Observing behaviour in English schools

Blatchford et al. (2003) conducted research into the impact of class size on pupil behaviour. They were interested in the possibility that, as class sizes increase, pupil inattentiveness also increases, resulting in difficult relationships between the children. The observational component of this research was based on children in large and small reception classes (age 4–5 years). The authors (2003: 21–2) describe their approach as involving

direct, i.e. on-the-spot, observations of selected children in terms of previously developed categories and in terms of 5-minute observation sheets divided into continuous 10-second time samples. The schedule was child-based in the sense that one child at a time was observed . . . The schedule involved categories that provided a description of time spent in three 'social modes'—when with their teachers, other children and when not interacting. Within each of these three modes sub-categories covered work, procedural, social and off-task activity. . . . The aim was to observe [six randomly chosen] children in each class five times per day, for 3 days. In the event the average number of completed observation sheets per child was 14 . . . In terms of time there were 69 minutes of observation per child.

The Blatchford et al. (2003) article emphasizes the data relating to interaction between children and 'off-task' behaviour. The interaction between children was observed and coded in terms of the following categories: task; procedure (class organization and management); social; mucking about; aggressive; help; and unclear.



The observation schedule

Devising a schedule for the recording of observations is clearly a crucial step in the structured observation project. The considerations that go into this phase are very similar to those involved in producing a structured interview schedule. The following considerations are worth taking into account.

- A clear focus is necessary. There are two aspects to this point. First, it should be clear to the observer exactly who or what (and possibly both) is to be observed. For example, if people are the focus of attention, the observer needs to know precisely who is to be observed. Also, the observer needs to know which if any aspects of the setting are to be observed and hence recorded. The second sense in which a clear focus is necessary is that the research problem needs to be clearly stated so that the observer knows which of the many things going on in any setting are to be recorded.
- As with the production of a closed question for a structured interview schedule or self-completion questionnaire, the forms taken by any category of behaviour must be both mutually exclusive (that is, not overlap) and inclusive. Taking the earlier example of coding behaviour in a university tutorial, we might conceivably run into a problem of the thirteen categories not being exhaustive if a student knocks on the tutor's door and quickly asks him or her a question (perhaps about the tutorial topic if the student is from another of the tutor's groups). An observer unfamiliar with the ways of university life might well be unsure about whether this behaviour needs to be coded in terms of the thirteen categories or whether the coding should be temporarily suspended. Perhaps the best approach would be to have another category of behaviour to be coded that we might term 'interruption'. It is often desirable for a certain amount of unstructured observation to take place before the construction of the observation schedule and for there to be some piloting of it, so that possible problems associated with a lack of inclusiveness can be anticipated.
- The recording system must be easy to operate. Complex systems with large numbers of types of behaviour will be undesirable. Much like interviewers using a structured interview schedule, observers need to be trained, but even so it is easy for an observer to become flustered or confused if faced with too many options.

- One possible problem with some observation schedules is that they sometimes require a certain amount of interpretation on the part of the observer. For example, it might be difficult to distinguish between a student responding to a question raised by another student and discussing the tutorial topic. To the extent that it may be difficult to distinguish between the two,

a certain amount of interpretation on the part of the observer may be required. If such interpretation is required, there would need to be clear guidelines for the observer, and considerable experience would be required (see Research in focus 12.2 for an illustration of a study in which a good deal of interpretation seems to have been necessary).



Research in focus 12.2

Observing jobs

Jenkins et al. (1975) report the results of an exploratory study employed to measure the nature of jobs. The research focused on several different types of job in a number of different types of organization. An observation schedule was devised to assess the nature of twenty aspects (dimensions) of the jobs in question. Most of the dimensions were measured through more than one indicator, each of which took the form of a question that observers had to answer on a six- or seven-point scale. These were then aggregated for each dimension. While the research has a predominantly psychological slant, many of the twenty dimensions relate to issues that have been raised in the sociology of work by labour process theorists and others (e.g. Braverman 1974). One dimension relates to 'Worker pace control' and comprises three observational indicators such as:

How much control does the employee have in setting the pace of his or her work?

Another dimension was 'Autonomy', which comprised four items, such as:

The job allows the individual to make a lot of decisions on his or her own.

Most of the observers were university students. The procedure for conducting the observations was as follows: 'Each respondent was observed twice for an hour. The observations were scheduled so that the two different observations were separated by at least 2 days, were usually made at different times of the day, and were always made by two different observers' (Jenkins et al. 1975: 173).



Strategies for observing behaviour

There are different ways of conceptualizing how behaviour should be recorded.

- We can record in terms of *incidents*. This means waiting for something to happen and then recording what follows from it. Essentially, this is what LaPiere (1934) did (see Thinking deeply 12.2), in that he waited for the Chinese couple to negotiate entry to each hotel or restaurant and then recorded whether they were allowed entry or not. I remember reading many years ago in a newspaper that someone placed a ladder over a pavement and then observed whether people preferred to go under the ladder or to risk life and limb in the face of oncoming traffic. A considerable number

preferred the latter option, confirming the persistence of superstitious beliefs in an apparently secular society. Once again, an incident (someone approaching the ladder) triggered the observation. Research in focus 12.3 contains a more recent example of such 'research'. Observation of this kind is what Webb et al. (1966) would regard as an example of *contrived observation*, because the researchers fabricated the situation. The discussion later in this chapter of *field stimulations* provides further illustrations of this kind of research.

- We can observe and record in terms of *short periods* of time. This was the case with the research reported in Research in focus 12.1, where '5-minute observation

sheets' were used. A slight variation on this theme can be found in the research reported in Research in focus 3.6. Children in St Helena were videotaped over a two-week period during their morning, lunchtime, and afternoon breaks. The tapes were then coded using

the Playground Behaviour Observation Schedule which is an instrument for recording the occurrence of 23 behaviours (e.g. games; fantasy play; character imitation; anti-social and pro-social behaviour) and their behaviour groupings (i.e. whether the behaviour was undertaken by an individual, a pair, or with 3–5 or 6 or more children). . . . A separate Playground Behaviour Observation Schedule was completed for each 30-second segment. (Charlton et al. 1998: 7)

- We can observe and record observations for quite *long periods* of time. The observer watches and records more or less continuously. The FIAC scheme adopts this strategy. Another example is the study of job characteristics by Jenkins et al. (1975), which entailed the observation of each worker on two occasions but for an hour on each occasion (see Research in focus 12.2):

'The observation hour was structured so that the observer spent 10 min becoming oriented to the job, 30 min observing specific job actions, and 20 min rating the job in situ. The observers then typically spent an additional 15 min away from the job completing the observation instrument' (Jenkins et al. 1975: 174). This last study is an example of what Martin and Bateson (1986) refer to as **continuous recording**, whereby the observer observes for extended periods, thus allowing the frequency and duration of forms of behaviour to be measured. They contrast this approach with **time sampling**.

- *Time sampling* is a further approach to the observation of behaviour. An example here would be a study of schools known as the ORACLE (Observational Research and Classroom Learning Evaluation) project (Galton et al. 1980). In this research, eight children (four of each gender) in each class in which observation took place were observed for around four minutes but on ten separate occasions. A mechanical device made a noise every twenty-five seconds, and, on each occasion this occurred, the observer made a note of what the teacher or pupils were doing in terms of the observation schedule. The sampling of time periods was random.



Research in focus 12.3

Is chivalry disappearing?

This is hardly research, but 'roadside tests' set up by the tyre firm Continental found that British male drivers were unwilling to help a 'woman described as blonde, in her mid-twenties and attractive' who was struggling to change a tyre (McVeigh 2006). These tests, which were a form of contrived observation, were set up in five British cities. Across the cities, 97 per cent of men zoomed past her, while she 'stood at the roadside clutching a spare wheel'. Chivalry was somewhat greater in Newcastle and among drivers of red cars! *The Times* columnist who reported this 'research' writes that a survey, for which no details are given, found that over half of British men claimed they would stop to help a woman struggling to change a tyre in such circumstances. This apparent gulf between actual and stated behaviour recalls the LaPiere (1934) study referred to in Thinking deeply 12.2.



Sampling

Just like survey research, structured observation necessitates decisions about sampling. However, with structured observation, issues surrounding sampling do not revolve solely around how to sample people. Several other sampling issues are involved.

Sampling people

When people are being sampled, considerations very similar to those encountered in Chapter 8 in respect of probability sampling come to the fore. This means that

the observer will ideally want to sample on a random basis. In Croll and Moses's (1985) research on children with special educational needs, thirty-four classrooms from a number of different schools were selected for observation. All children were within the same age range. Initially, each teacher was interviewed to determine which children in his or her class were regarded as having special needs. In addition, tests of both reading ability and non-verbal reasoning were administered to children to identify those who appeared to have special needs, but who had not been identified by the teacher. Up to six children with special needs in each class were then randomly sampled as being the focus of structured observation; so too were four children who had not been identified as having special needs—these children acted as a kind of control group. In this way, 280 children were sampled, of whom 151 were identified as having special needs; the other 129 served as control subjects. The teachers did not know exactly which children were being observed. Each child was observed for a few minutes, after which the observer proceeded to the next child to be observed in a predetermined random order. In the end, each child was observed for a total of two hours. This was made up of a large number of short observation periods. In the research reported in *Research in focus 12.1*, the six students in each class were selected randomly but with the stipulation that three boys and three girls would be sampled. In the study of job characteristics (*Research in focus 12.2*), the individuals who were observed at work were randomly selected (Jenkins et al. 1975).

Sampling in terms of time

As implied by the idea of time sampling (see above), it is often necessary to ensure that, if certain individuals are sampled on more than one occasion, they are not always observed at the same time of the day. This means that, if particular individuals are selected randomly for observation on several different occasions for short periods, it is desirable for the observation periods to be randomly selected. For example, it would not be desirable for a certain pupil in a school classroom always to be observed at the end of the day. He or she may be tired, and this will give a false impression of that pupil's behaviour. In the research reported in *Research in focus 12.1*, each child was observed at separate times on three different days. As a result, the researchers' ratings of any child are unlikely to be distorted by unusual behaviour that he or she might exhibit on just one or two occasions.

Further sampling considerations

The sampling procedures mentioned so far conform to probability sampling principles, because it is feasible to construct a sampling frame for individuals. However, this is not always possible for different kinds of reason. Studies in public areas, like the research on superstition mentioned above, do not permit random sampling, because we cannot very easily construct a sampling frame of people walking along a street. Similarly, it is not feasible to construct a sampling frame of interactions. Reiss (1976), for example, has written about the difficulty of developing a random sample of encounters between police officers and the public. The problem with doing structured observation research on such a topic is that it does not lend itself to the specification of a sampling frame, and therefore the researcher's ability to generate a probability sample is curtailed.

As suggested in Chapter 8, considerations relating to probability sampling derive largely from concerns surrounding the external validity (or generalizability) of findings. Such concerns are not necessarily totally addressed by resorting to probability sampling, however. For example, if a structured observation study is conducted over a relatively short span of time, issues of the representativeness of findings are likely to arise. If the research was conducted in schools, observations conducted towards the end of the school year, when examinations are likely to loom large in the thinking of both teachers and students, may affect the results obtained compared to observations at a different point in the academic year. Consequently, consideration has to be given to the question of the timing of observation. This potential problem was dealt with in the ORACLE research by ensuring that teachers and each target pupil were observed on six occasions during each of the three school terms. Furthermore, how are the sites in which structured observation is to take place selected? Can we presume that they are themselves representative? Clearly, a random sampling procedure for the selection of schools may assuage any worries in this connection. However, in view of the difficulty of securing access to settings such as schools and business organizations, it is likely that the organizations to which access is secured may not be representative of the population of appropriate ones.

A further set of distinctions between types of sampling in structured observation have been drawn by Martin and Bateson (1986) between:

- **ad libitum sampling**, whereby the observer records whatever is happening at the time;

- **focal sampling**, in which a specific individual is observed for a set period of time; the observer records all examples of whatever forms of behaviour are of interest in terms of a schedule;
- **scan sampling**, whereby an entire group of individuals is scanned at regular intervals and the behaviour of all of them is recorded at that time. This sampling strategy allows only one or two types of behaviour to be observed and recorded; and

- **behaviour sampling**, whereby an entire group is watched and the observer records who was involved in a particular kind of behaviour.

Most structured observation research seems to employ the first two types: Flanders's FIAC scheme is an example of *ad libitum* sampling; the research by Galton et al. (1980), Croll and Moses (1985), Blatchford et al. (2003; see Research in focus 12.1), Jenkins et al. (1975; see Research in focus 12.2), and the research by Buckle and Farrington (1994) cited in Research in focus 12.4 are illustrations of focal sampling.



Research in focus 12.4

A study of shoplifting

Buckle and Farrington (1994) report the results of a replication of an earlier study of shoplifting in a department store in Peterborough (Buckle and Farrington 1984). The replication was conducted in a similar store in Bedford. Customers were selected at random as they entered the store and followed by two observers until they left. The observers recorded such details as: cost of items bought; gender, race, and estimated age; and behaviour. In Peterborough 486 people formed the basis of the sample and in Bedford it was 502. Nine people shoplifted in Peterborough and six in Bedford. Somewhat surprisingly, shoplifters were more likely to be male and either under 25 (in Peterborough) or over 55 (in Bedford). Most shoplifters also purchased goods. Most shoplifting was of small items of relatively little monetary value. The sampling and observation strategies entailed random sampling of people followed by continuous recording for a short or long period depending on how long the person remained in the store.



Issues of reliability and validity

One writer has concluded that, when compared to interviews and questionnaires, structured observation 'Provides (a) more reliable information about events; (b) greater precision regarding their timing, duration, and frequency; (c) greater accuracy in the time ordering of variables; and (d) more accurate and economical reconstructions of large-scale social episodes' (McCall 1984: 277). This is a very strong endorsement for structured observation, but, as McCall notes, there are several issues of reliability and validity that confront practitioners of the method. Some of these issues are similar to those faced by researchers when seeking to develop measures in social research in general (see Chapter 7) and in survey research in particular. However, certain concerns are specific to structured observation.

Reliability

Practitioners of structured observation have been concerned with the degree of *inter-observer consistency*. Essentially, this issue entails considering the degree to which two or more observers of the same behaviour agree in terms of their coding of that behaviour on the observation schedule. The chief mechanism for assessing this component of reliability is a statistic called *kappa* (see Key concept 12.3; *this box can be ignored if you feel unsure about addressing more complex statistical issues at this stage*).

A second consideration in relation to reliability is the degree of consistency in the application of the observation schedule over time—that is, *intra-observer consistency*.

This is clearly a difficult notion, because of the capacity and often necessity for people to behave in different ways on different occasions and in different contexts. Assessing the consistency of observation ratings across all possibilities is clearly a difficult undertaking. The procedures for assessing this aspect of reliability are broadly similar to those applied to the issue of inter-observer consistency. The Jenkins et al. (1975) research addressed the issue of inter-observer consistency over time and found that the measures fared even worse in this respect (Research in focus 12.2).

It is clearly not an easy matter to achieve reliability in structured observation. This is a point of some significance in view of the fact that validity presupposes reliability (see Chapter 7). Reliability may be difficult to achieve on occasions, because of the effects of such

factors as observer fatigue and lapses in attention. However, this point should not be exaggerated, because the ORACLE researchers were able to achieve high levels of reliability for many of their measures, and indeed two critics of structured observation have written that 'there is no doubt that observers can be trained to use complex coding schedules with considerable reliability' (Delamont and Hamilton 1984: 32). (Using the Scott coefficient referred to in Key concept 12.3, it was found that the average inter-observer reliability level across the different components of pupil behaviour was 0.90.) The high levels of reliability may be due to the factors that are intrinsic to the school classroom and to the fact that there is a long tradition of structured observation research in schools, so that a fund of experience has been accumulated in this domain.



Key concept 12.3 Cohen's kappa

This box can be ignored if you feel unsure about addressing more complex statistical issues at this stage.

Cohen's kappa is a measure of the degree of agreement over the coding of items by two people. As such, it could be applied to the coding of any textual information, as in the content analysis of newspaper articles or of answers to open interview questions, as well as to the coding of observation. Much like Cronbach's alpha (see Key concept 7.4) you will end up with a coefficient that will vary between 0 and 1. The closer the coefficient is to 1, the higher the agreement and the better the inter-observer consistency. A coefficient of 0.75 or above is considered very good; between 0.6 and 0.75, it is considered good; and between 0.4 and 0.6, it is regarded as fair. The meaning of kappa is that it measures the degree of agreement between observers beyond that which would occur by chance. Croll (1986) refers to a very similar statistic, the Scott coefficient of agreement, which can be interpreted in an identical way.

The values of kappa in the study of job characteristics referred to in Research in focus 12.2 were mainly in the 'fair' category. The two items referred to in Research in focus 12.2 achieved kappa values of 0.43 and 0.54 respectively (Jenkins et al. 1975). These are not very encouraging and suggest that the coding of job characteristics was not very reliable. By contrast, the kappa coefficients for the various measures used in the observational research on pupils' engagement in class by Blatchford et al. (2009) that is mentioned elsewhere in this chapter ranged between 0.80 and 0.77, which indicates a high level of inter-observer consistency.

Validity

Measurement validity relates to the question of whether a measure is measuring what it is supposed to measure. The validity of any measure will be affected by:

- whether the measure reflects the concept it has been designed to measure (see Chapter 7), and

- error that arises from the implementation of the measure in the research process (see Chapter 9).

The first of these issues simply means that in structured observation it is necessary to attend to the same kinds of issues concerning the checking of validity (assessing face validity, concurrent validity, and so on) that are encountered in research based on interviews and

questionnaires. The second aspect of validity—error in implementation—relates to two matters in particular.

- Is the observation schedule administered as it is supposed to be? This is the equivalent of ensuring that interviewers using a structured interview schedule follow the research instrument and its instructions exactly as they are supposed to. If there is variability between observers or over time, the measure will be unreliable and therefore cannot be valid. Ensuring that observers have as complete an understanding as possible of how the observation schedule should be implemented is therefore crucial.
- Do people change their behaviour because they know they are being observed? This is an instance of what is known as the ‘reactive effect’ (Key concept 12.4). After all, if people adjust the way they behave because they know they are being observed (perhaps because they want to be viewed in a favourable way by the observer), their behaviour would have to be considered atypical. As a result, we could hardly regard the results of structured observation research as indicative of what happens in reality. As McCall (1984) notes, there is evidence that a reactive effect occurs in structured observation, but that by and large research participants become accustomed to being observed, so that the researcher essentially becomes less intrusive the longer he or she is present. Moreover, it should be borne in mind that frequently people’s awareness of the observer’s presence is offset by other factors. For example, teachers and students have many tasks to accomplish that reflect

the demands of the classroom, so that the observer’s ability to make a big impact on behaviour may be curtailed by the requirements of the situation. A study by Harvey et al. (2009) that entailed an evaluation of a programme to promote the use of insecticide-treated bed nets to reduce the incidence of malaria in the Peruvian Amazon illustrates this point. Four observers monitored sixty households in four rural villages continuously for twelve-hour periods. The observers informed the villagers they were observing that they must not interact with each other so that reactivity could be reduced. The observers recorded 339 instances of reactivity. At their most basic, these entailed salutations to the observer or interaction with the observer over matters unrelated to malaria prevention. On only two occasions was there evidence of reactivity that related to the study objectives. On these two occasions it was clear that the villagers changed their behaviour as a result of the observer’s presence. Here is an example from the field notes of the observer concerned:

18:20: I put on my long-sleeved shirt because there are a lot of mosquitoes. The 33-year-old man enters his room and I see him return with his own long-sleeve shirt. I think he put it on because he saw me put on mine. (Harvey et al. 2009: 14)

These findings are quite reassuring in suggesting that the reactive effect of observation is not as problematic as might be anticipated.



Key concept 12.4

Reactive effect

Webb et al. (1966: 13) wrote about the ‘reactive measurement effect’, by which they meant that ‘the research subject’s knowledge that he is participating in a scholarly search may confound the investigator’s data’. They distinguished four components of this effect.

1. *The guinea pig effect—awareness of being tested.* Examples of the kind of concern that Webb et al. were writing about are such effects as the research participant wanting to create a good impression or feeling prompted to behave in ways (or express attitudes) that would not normally be exhibited.
2. *Role selection.* Webb et al. argue that participants are often tempted to adopt a particular kind of role in research. An example is that there is a well-known effect in experimental research (but which may have a broader applicability) whereby some individuals seek out cues about the aims of the research and adjust what they say and do in line with their perceptions (which may of course be false) of those aims.
3. *Measurement as a change agent.* The very fact of a researcher being in a context in which no researcher is normally present may itself cause things to be different. For example, the fact that there is an observer sitting

in the corner of a school classroom means that there is space and a chair being used that otherwise would be unoccupied. This very fact may influence behaviour.

4. *Response sets.* This is an issue that primarily relates to questionnaire and interview research and occurs when the respondent replies to a set of questions in a consistent but clearly inappropriate manner. Examples of this kind of effect are measurement problems like the social desirability effect and yeasaying and naysaying (consistently answering yes or no to questions or consistently agreeing or disagreeing with items regardless of the meaning of the question or item).

Reactive effects are likely to occur in any research in which participants know they are the focus of investigation. Webb et al. called for greater use of what they call *unobtrusive measures* or *non-reactive methods* that do not entail participants' knowledge of their involvement in research (see Key concept 14.3 for more information).



Field stimulations as a form of structured observation

Salancik (1979) has used the term **field stimulation** to describe a form of observation research that shares many of the characteristics of structured observation. Although he classifies field stimulations as a qualitative method, they are in fact better thought of as operating with a quantitative research strategy, since the researcher typically seeks to quantify the outcomes of his or her interventions. In terms of the classification offered in Key concept 12.2, it is in fact 'contrived observation'. Part of LaPiere's (1934) study (see Thinking deeply 12.2) was a field stimu-

lation: when he arranged for the Chinese couple to seek entry to the hotels and restaurants in order to observe the effects of their attempts, he was employing a field stimulation. A field stimulation, therefore, is a study in which the researcher directly intervenes in and/or manipulates a natural setting in order to observe what happens as a consequence of that intervention. However, unlike most structured observation, in a field stimulation participants do not know they are being studied. A famous field stimulation is described in Research in focus 12.5.



Research in focus 12.5 A field stimulation

David Rosenhan (1973) was one of eight people who sought to gain entry as patients to mental hospitals in the USA. Some of them—they are referred to as 'pseudo-patients'—sought entry to more than one hospital, so that twelve hospitals were approached. Each pseudo-patient was instructed to say that he or she was hearing voices. All successfully gained entry, in eleven of the twelve cases with a diagnosis of schizophrenia. As soon as they had succeeded in gaining entry, the pseudo-patients were instructed to cease exhibiting any symptoms. In spite of the fact that the pseudo-patients were all 'sane', it took many of them quite a long time to be released. The length of hospitalization varied between seven and fifty-two days with a mean of nineteen days. In four of the hospitals, pseudo-patients approached psychiatrists and nurses with a request for release, with no member of staff being approached more than once on any day. The pseudo-patients recorded the nature of the response to their requests: 71 per cent of psychiatrists responded by moving on with their heads averted and 88 per cent of nurses did likewise. Rosenhan regards this evidence as indicating that the mental patient becomes powerless and depersonalized. The study has been highly controversial, as many psychiatrists have sought to question its implications, while others have raised ethical issues of the kind addressed in Chapter 6 (such as the use of deception).

Table 12.1

Daniel's (1968) situation test: the case of accommodation

Reaction to request for accommodation	No.
West Indian was told accommodation taken; both other applicants told it was vacant	38
West Indian was asked for higher rent than the others	4
West Indian and Hungarian were told accommodation let	2
West Indian and Hungarian were asked for higher rent	1
All applicants received the same information	15
TOTAL	60

Some field stimulations can take the form of an experimental design (see Chapter 3). An example is a study by Daniel (1968) of racial discrimination in Britain in the 1960s. Daniel undertook conventional attitude studies among immigrant groups to establish levels of discrimination. In addition, he developed 'situation tests' to back up his findings. For example, in one set of situation tests he examined discrimination in the area of accommodation. Sixty advertisements for accommodation to let were selected from a number of regions. Advertisements stipulating 'no coloureds' or 'Europeans only' were deliberately excluded. At the time, it was not illegal for landlords to place such instructions in their advertisements. Each landlord was approached by each of the following: a West Indian; a white Hungarian; and a white Englishman. The applicants were presented with identical sets of characteristics, but they differed in terms of ethnicity. The applicant was requesting accommodation for a married couple with no children. In half of the applications (that is, thirty), the testers adopted 'professional roles'. In these roles they sought more expensive accommodation. In the other half, they adopted manual roles. In fifteen of the sixty cases, all three applicants got the same information (for example, let, still vacant). This means that discrimination occurred in the remaining forty-five cases (see Table 12.1).

Daniel's research strongly suggests that, because the Hungarian was rarely discriminated against, it is colour rather than being a member of an ethnic minority as such that causes discrimination. Similar studies were conducted in relation to house purchase, employment, and car insurance. Interestingly, the researchers often found that these tests implied that discrimination was *greater* than had been indicated by the attitude surveys, presumably because it is difficult to know if you really have been discriminated against.

While such research provides some quite striking findings and gets around the problem of **reactivity** by not alerting research participants to the fact that they are being observed, like the pseudo-patient study in Research in focus 12.5, ethical concerns are sometimes raised, such as the use of deception. Moreover, the extent to which an observation schedule can be employed is inevitably limited, because excessive use will blow the observer's cover. All that can usually be done is to engage in limited coding, in particular the nature of the effect of the intervention, as in the LaPiere (1934) and Daniel (1968) studies, or to include a limited amount of further observation, as in the Rosenhan (1973) research.



Criticisms of structured observation

Although it is not extensively used in social research, structured observation has been quite controversial. Certain criticisms have been implied in some of the previous discussion of reliability and validity issues, as well as in connection with the issue of generalizability.

However, certain other areas of criticism warrant further discussion.

- There is a risk of imposing a potentially inappropriate or irrelevant framework on the setting being observed.

This point is similar to the problem of the closed question in questionnaires. This risk is especially great if the setting is one about which little is known. One solution is for the structured observation to be preceded by a period of unstructured observation, so that appropriate variables and categories can be specified.

- Because it concentrates upon directly observable behaviour, structured observation is rarely able to get at intentions behind behaviour. Sometimes, when intentions are of concern, they are imputed by observers. Thus, in the FIAC scheme (see Figure 12.1), the category 'teacher praises or encourages' means imputing a motive to something that the teacher says. Similarly, Blatchford et al. (2009: 668) report that one of the categories of observation of pupil behaviour they employed in their study of the impact of teaching assistants on engagement in class was: '*Individual off-task (passive): target child is disengaged during task activity, for example, day dreaming.*' Essentially, the problem is that structured observation does not readily allow the observer to get a grasp of the meaning of behaviour.
- There is a tendency for structured observation to generate lots of fragments of data. The problem here can be one of trying to piece them together to produce an overall picture, or one of trying to find general themes that link the fragments of data together. It becomes difficult, in other words, to see a bigger picture that lies behind the segments of behaviour that structured observation typically uncovers. It has been suggested, for example, that the tendency for structured observation studies of managers at work to find little evidence of planning in their everyday work (e.g. Mintzberg 1973) is due to the tendency for the method to fragment a manager's activities into discrete parts. As a result, something like planning, which may be an element in many managerial activities, becomes obscured from view (Snyder and Glueck 1980).
- It is often suggested that structured observation neglects the context within which behaviour takes place. Delamont and Hamilton (1984), for example, note in connection with the ORACLE research that it was found that teachers' styles were related to their ages. However, they argue that such a finding can really be understood only 'if data are gathered on teacher careers and life histories of a kind eschewed by ORACLE' (1984: 9). Of course, were such data collected, this criticism would have little weight, but the tendency of structured observation researchers to concentrate on overt behaviour tends to engender this kind of criticism.

On the other hand . . .

It is clear from the previous section that there are undeniable limitations to structured observation. However, it also has to be remembered that, when overt behaviour is the focus of analysis and perhaps issues of meaning are less salient, structured observation is almost certainly more accurate and effective than getting people to report on their behaviour through questionnaires. Also, although the point was made in the previous section that the observation of behaviour often necessitates imputing meaning to it, that is not to say that imputing meaning is *always* involved. With most of the categories of behaviour used by Blatchford et al. (2009), little of any assignment of motive is required. Also, if video evidence is accumulated, the researcher is afforded the opportunity to review the evidence at length and not rush to a possibly snap decision about what is being observed. For example, Sampson and Raudenbush (1999) took video footage of Chicago streets to develop a measure of social disorder that included such indicators as: alcohol consumption in public; sale of drugs; street prostitution; and fights between adults or hostile arguing.

It may also be that structured observation is a method that works best when accompanied by other methods. Since it can rarely provide reasons for observed patterns of behaviour, if it is accompanied by another method that can probe reasons, it is of greater utility. Delamont (1976) in her research in a school found FIAC to be useful as a means of exploring differences in teaching style between teachers. However, she was able to get at some of the reasons for the quantitative differences that she discerned only because she had carried out some participant observation and semi-structured interviewing (two of the main methods of qualitative research) in various school classes. For example, she compared two Latin teachers who were similar in certain respects but differed in terms of 'the proportion of questioning to lecturing in their speech' (Delamont 1976: 108). These differences in teaching style reflected contrasting views about teaching and differences in personal demeanour. Blatchford (2005) reports that the structured observation data that were collected in the research reported in Research in focus 12.1 were part of a wider study of the impact of variations in class size. The other methods employed were: termly questionnaires administered to teachers to gauge their estimates of how they allocated time in classrooms between different activities; end-of-year questionnaires administered to teachers asking them about their experiences of the impact of class size; and case studies of small and large classes comprising some semi-structured

observation of events and semi-structured interviews with teachers and the head teacher.

In laboratory experiments in fields like social psychology, observation with varying degrees of structure is quite commonplace, but in social research structured observation has not been frequently used. Perhaps one major reason is that, although interviews and questionnaires are limited in terms of their capacity to tap behaviour accurately, as noted above, they do offer the opportunity to reveal information about both behaviour *and* attitudes and social backgrounds. In other words, they are more flexible and offer the prospect of being able to uncover a variety of correlates of behaviour (albeit

reported behaviour), such as social background factors. They can also ask questions about attitudes and investigate explanations that people proffer for their behaviour. As a result, researchers using questionnaires are able to gain information about some factors that may lie behind the patterns of behaviour they uncover. Also, not all forms of behaviour are liable to be accessible to structured observation and it is likely that survey research or researcher-driven diaries (see Key concept 10.1) are the only likely means of gaining access to them. However, greater use of structured observation may result in greater facility with the method, so that reliable measures of the kind developed in areas like education might emerge.



Checklist

Structured observation research

- ☐ Have you clearly defined your research questions?
- ☐ Is the sample to be observed relevant to your research questions?
- ☐ Can you justify your sampling approach?
- ☐ Does your observation schedule indicate precisely which kinds of behaviour are to be observed?
- ☐ Have your observation categories been designed so that there is no need for the observer to interpret what is going on?
- ☐ Have you made sure that the categories of behaviour do not overlap?
- ☐ Do all the different categories of behaviour allow you to answer your research questions?
- ☐ Have you piloted your observation schedule?
- ☐ Are the coding instructions clear?
- ☐ Are the categories of behaviour inclusive?
- ☐ Is it easy to log the behaviour as it is happening?



Key points

- Structured observation is an approach to the study of behaviour that is an alternative to survey-based measures.
- It comprises explicit rules for the recording of behaviour.
- Structured observation has tended to be used in relation to a rather narrow range of forms of behaviour, such as that occurring in schools.
- It shares with survey research many common problems concerning reliability, validity, and generalizability.

- Reactive effects have to be taken into account but should not be exaggerated.
- Field stimulations represent a form of structured observation but suffer from difficulties concerning ethics.
- Problems with structured observation revolve around the difficulty of imputing meaning and ensuring that a relevant framework for recording behaviour is being employed.



Questions for review

Problems with survey research on social behaviour

- What are the chief limitations of survey research with regard to the study of behaviour?

So why not observe behaviour?

- What are the chief characteristics of structured observation?
- To what extent does it provide a superior approach to the study of behaviour than questionnaires or structured interviews?

The observation schedule

- What is an observation schedule?
- 'An observation schedule is much like a self-completion questionnaire or structured interview except that it does not entail asking questions.' Discuss.
- Devise an observation schedule of your own for observing an area of social interaction in which you are regularly involved. Ask people with whom you normally interact in those situations how well they think it fits what goes on. Have you missed anything out?

Strategies for observing behaviour

- What are the main ways in which behaviour can be recorded in structured observation?

Sampling

- Identify some of the main sampling strategies in structured observation.

Issues of reliability and validity

- How far do considerations of reliability and validity in structured observation mirror those encountered in relation to the asking of questions in structured interviews and self-completion questionnaires?
- What is the reactive effect and why might it be important in relation to structured observation research?

Field stimulations as a form of structured observation

- What are field stimulations and what ethical concerns are posed by them?

Criticisms of structured observation

- 'The chief problem with structured observation is that it does not allow us access to the intentions that lie behind behaviour.' Discuss.
- How far do you agree with the view that structured observation works best when used in conjunction with other research methods?